Multiple Instance Regression with Structured Data

Kiri L. Wagstaff (Jet Propulsion Lab., California Inst. of Tech.),
Terran Lane (University of New Mexico),
and Alex Roper (California Institute of Technology)

Workshop on Mining Complex Data
December 15, 2008

THE UNIVERSITY of NEW MEXICO

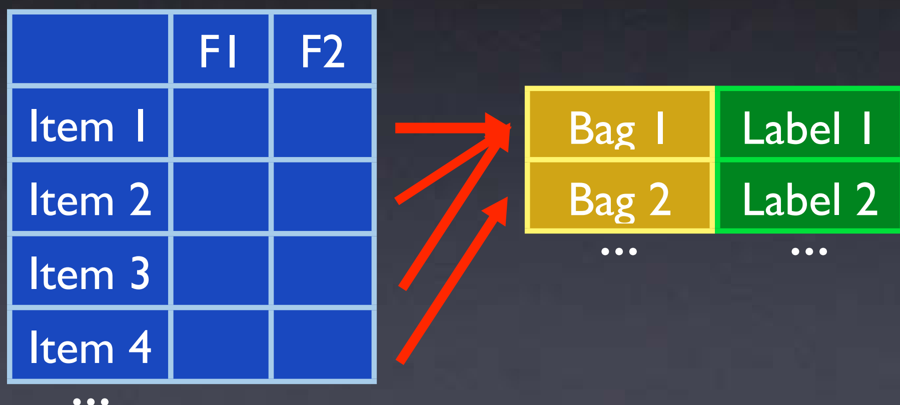
Government sponsorship acknowledged.

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

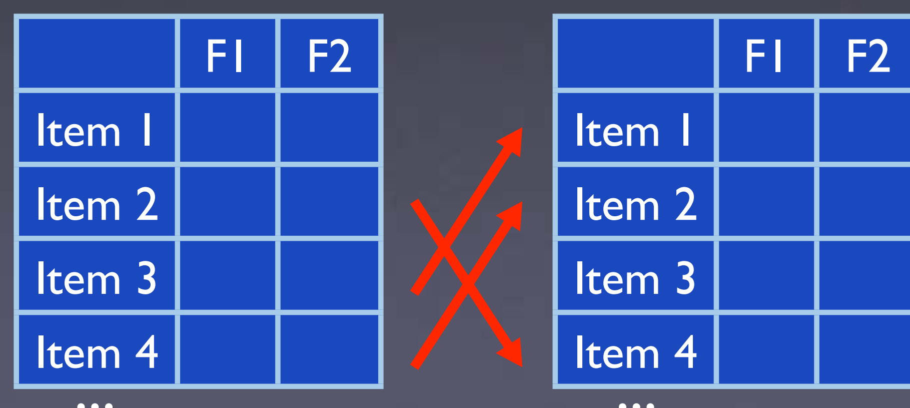
Data Structure Spectrum

Tabular Data

Multiple-Instance Data

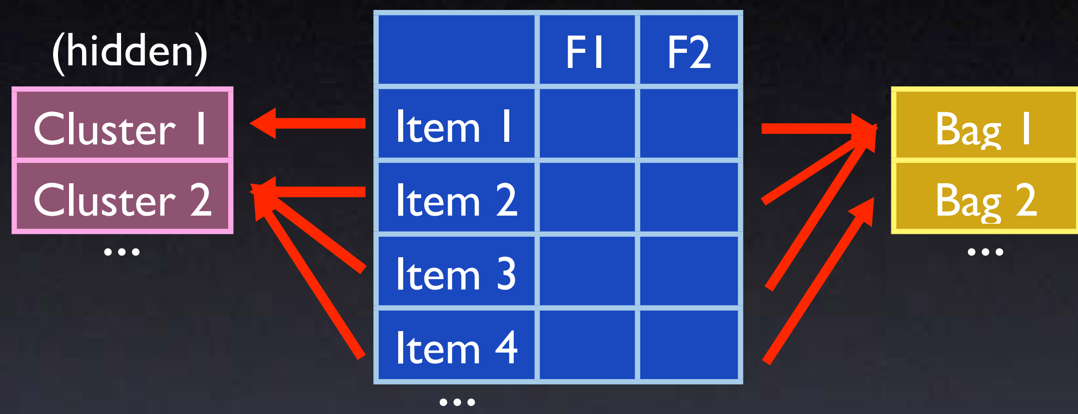


Relational Data



MIL with Structured Bags

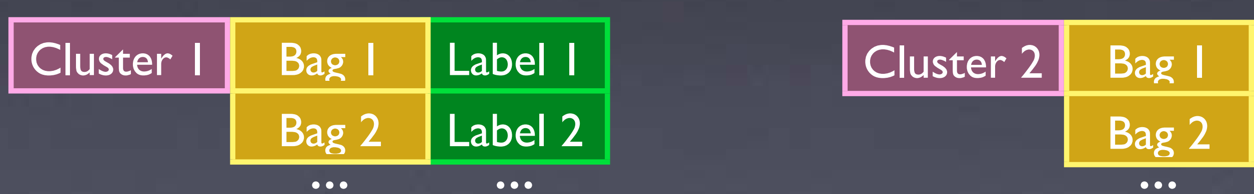
Bags contain sub-populations (clusters)



Only one cluster is relevant to the target concept

Relevant

Irrelevant



Items contribute to bag labels only through cluster membership

Structure =

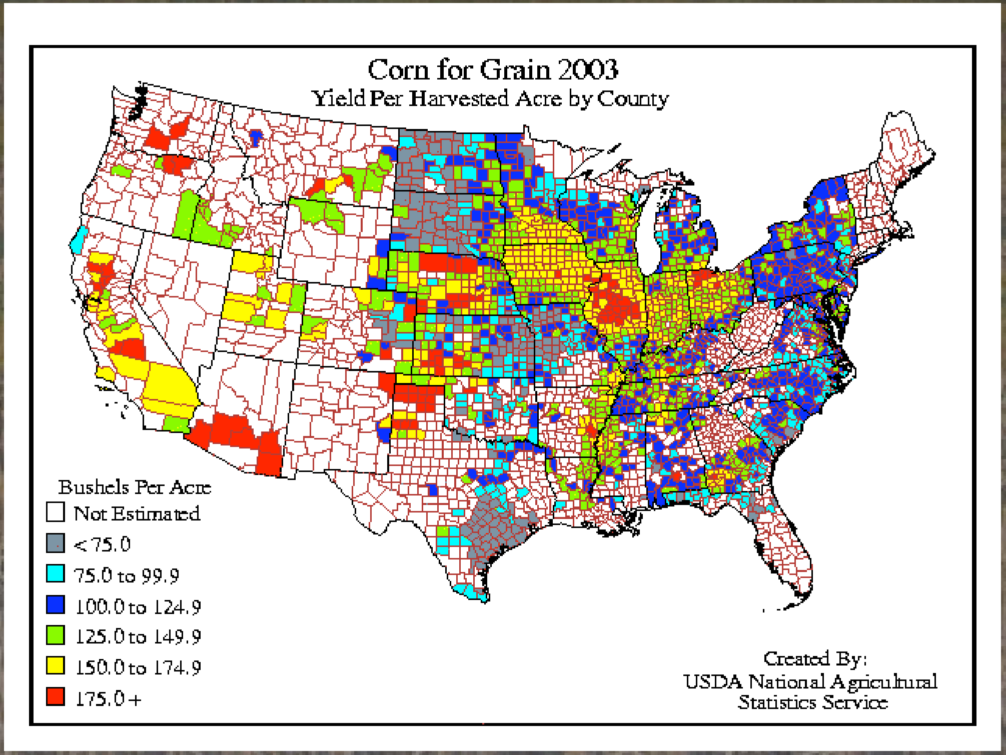
bag contents drawn from multiple distributions

What problems have this structure?

Predicting Crop Yield

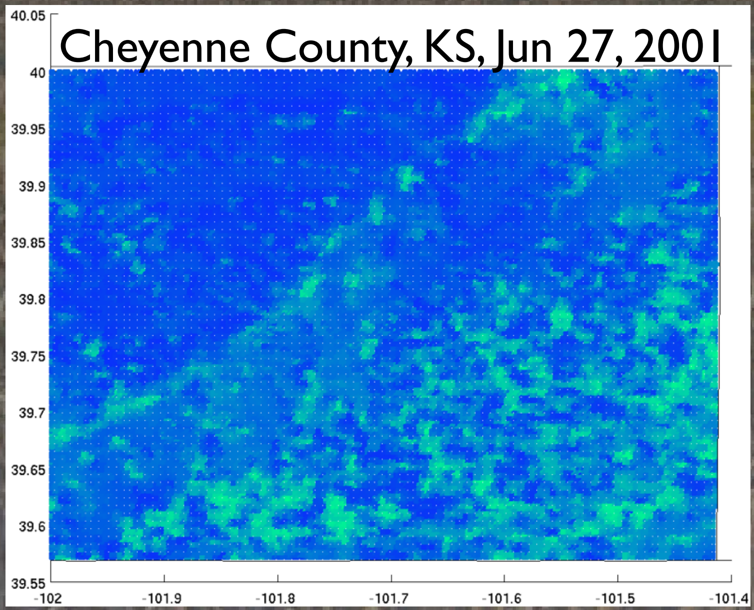
- USDA:
  - Post-harvest yield results per county, per crop
  - Could we predict yield earlier in the year?
  - Data = remote sensing: weekly observations, entire U.S.

- Benefits:
  - Inform agricultural markets
  - Enable more focused precision agriculture



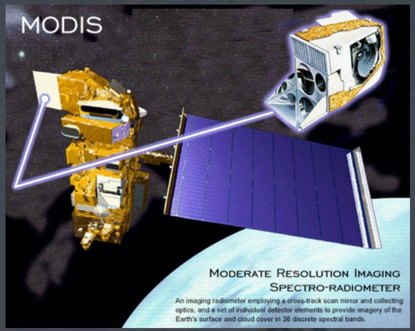
Multiple Instance Problem

- Each county (bag of pixels):
  - 250 m/pixel = 30,000 - 300,000 pixels
  - One label per crop: bushels/acre
- Which ones are relevant?
  - Bags have structure
  - Sub-pixel mixing: Need to model degree of membership




```
37 bu/acre of wheat
124 bu/acre of corn
```

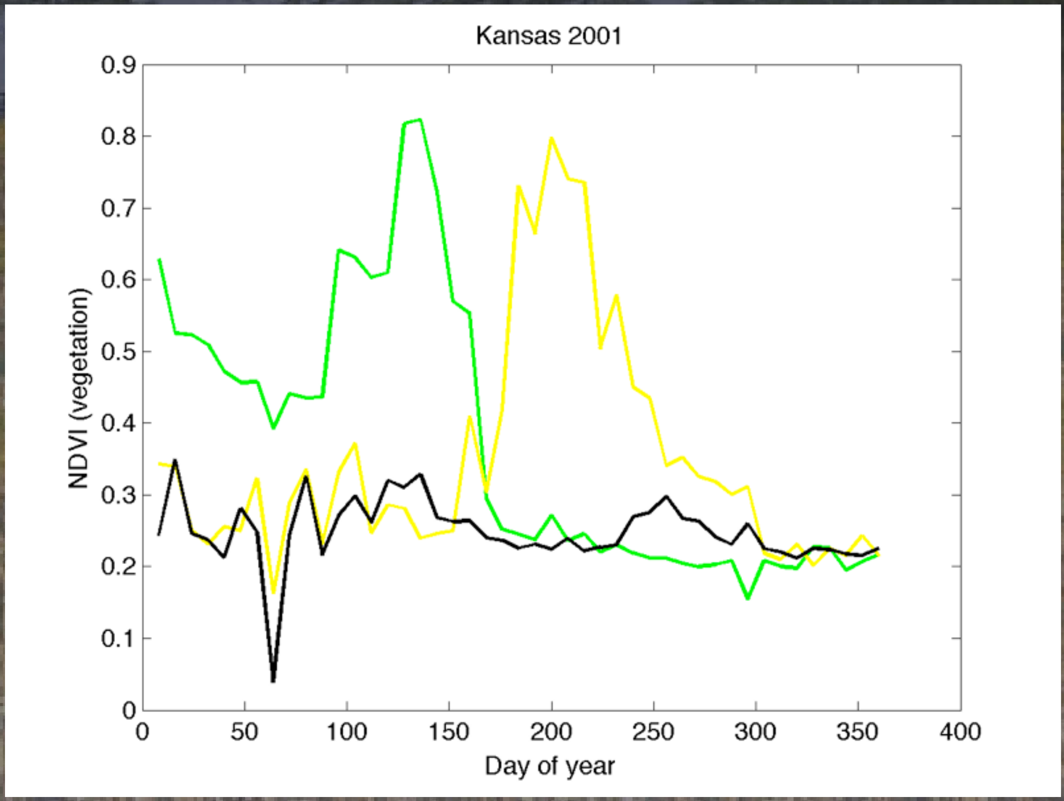
Instance = Time Series



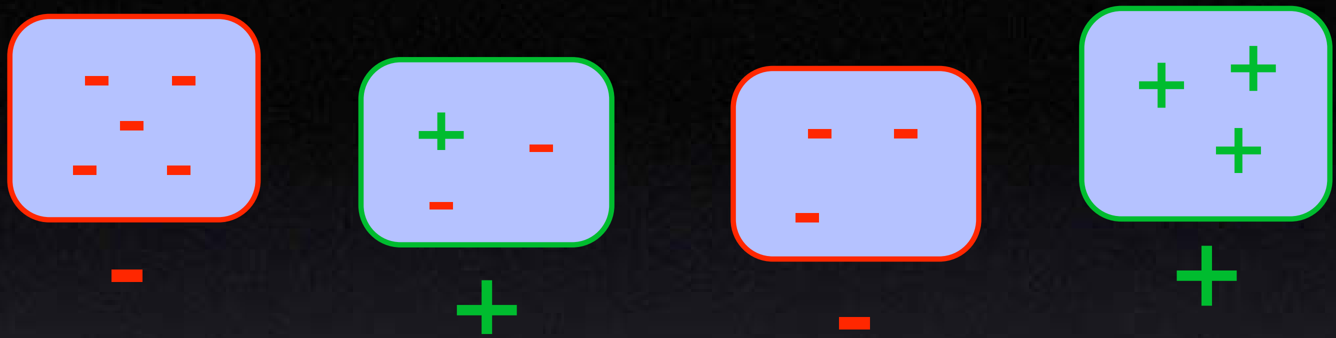
- MODIS: Red and NIR every 8 days

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$$

- How early can we make good predictions?
- Time series can reveal crop type
  - Or at least crop vs. forest/city/etc.
  - Thus hinting at relevance to label



Multiple-Instance Learning



- Classification: >= 1 positive item -> positive bag [Dietterich et al., 97]
- MIL via Embedded Instance Selection (MILES) [Chen et al., 06]
  - Embed bags in item-similarity feature space,
    use feature selection to find relevant ones, use regular SVM
  - Application: region-based image categorization

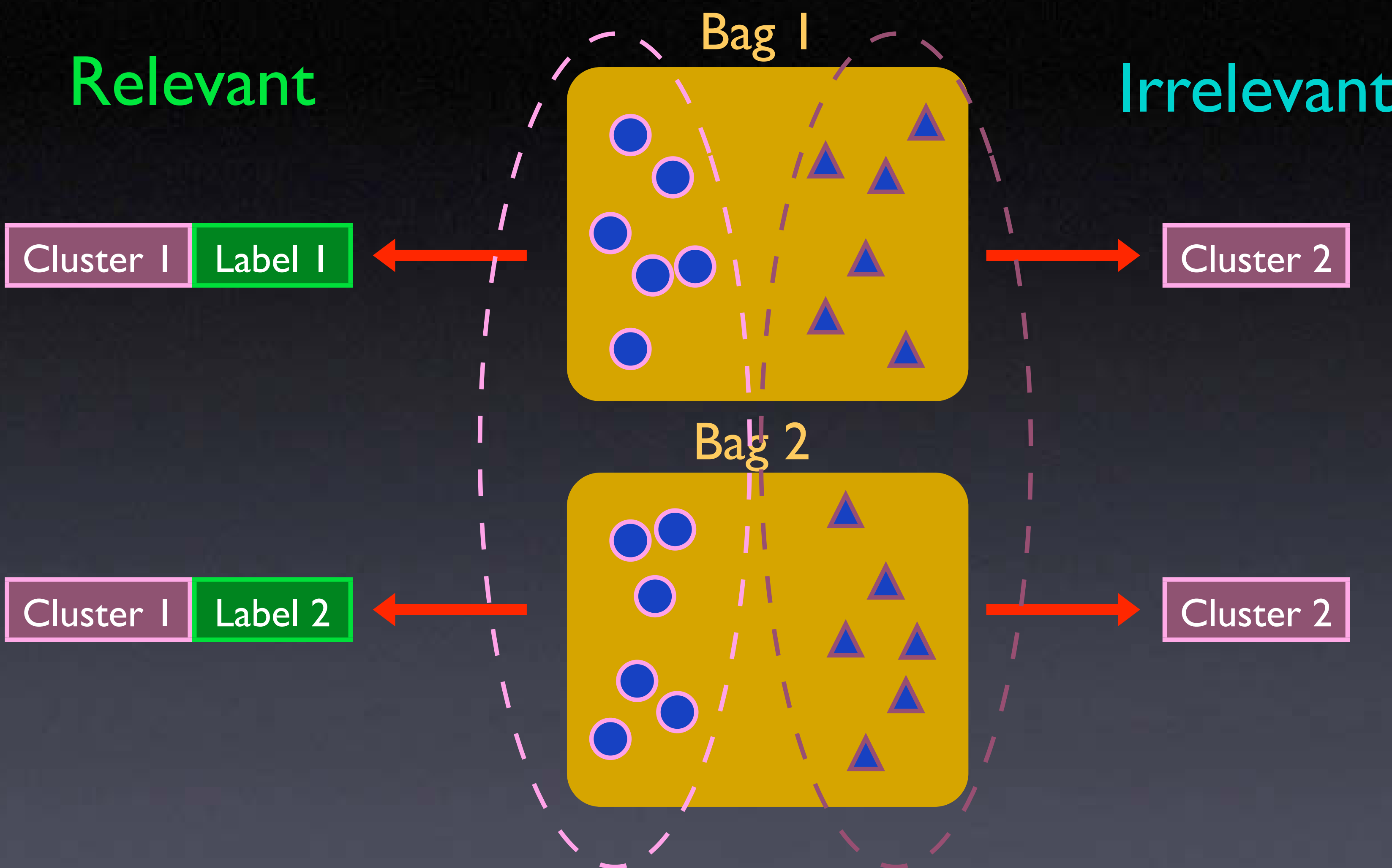
Multiple-Instance Regression

- Primary Instance Regression (PIR) [Ray & Page, 01]
  - Find single item that dictates bag label
  - Other items are noisy observations of primary

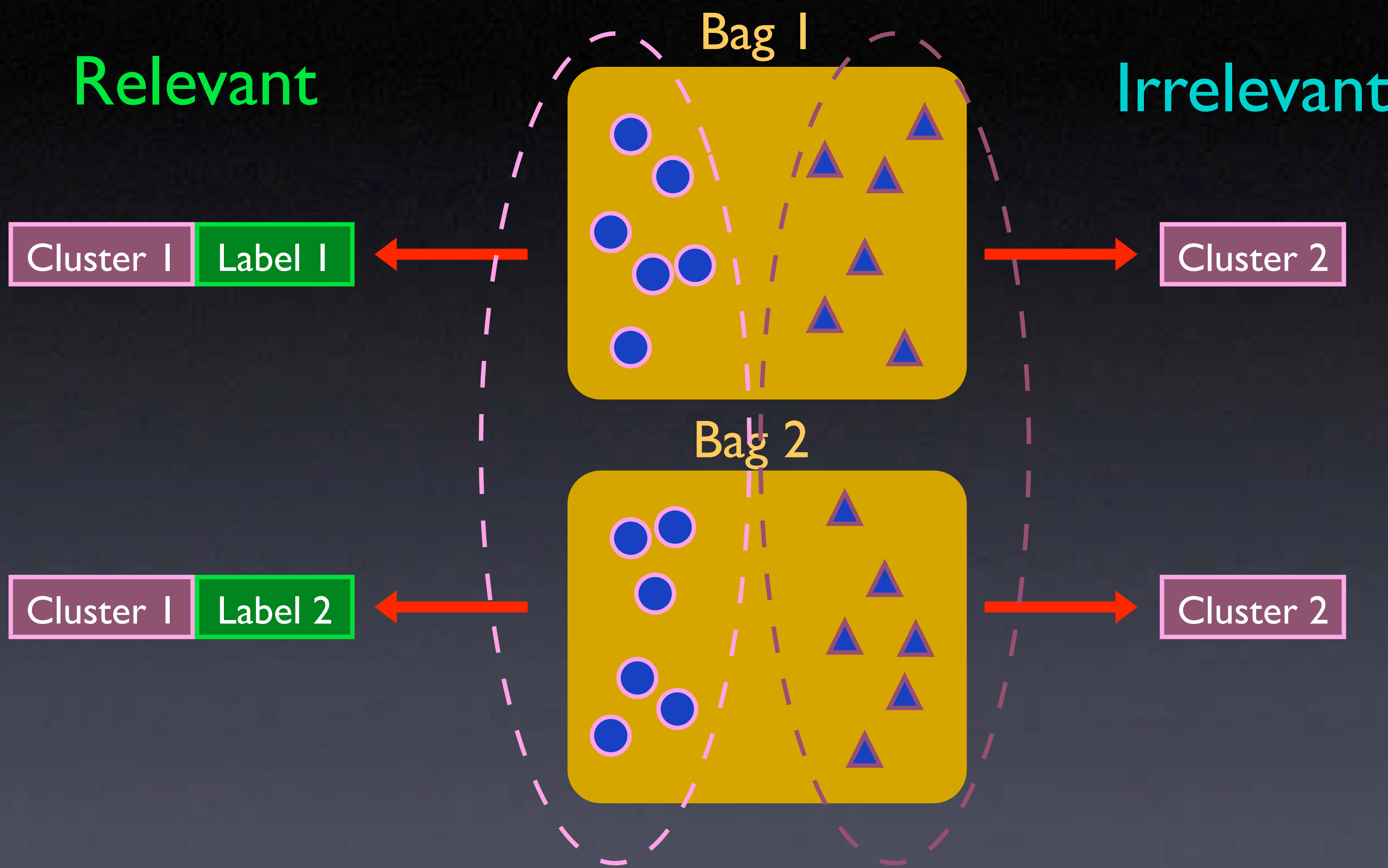
Our Solution: Cluster Regression Models

- Explicitly model bag structure, multiple populations
- Assumption: bag label derives from a subset of *similar* items (in input feature space)
  - Individual relevance per item
- Approach:
  1. Identify clusters of items
  2. Build one regression model per cluster
  3. Select model that best fits the bag labels

MIL with Structured Bags

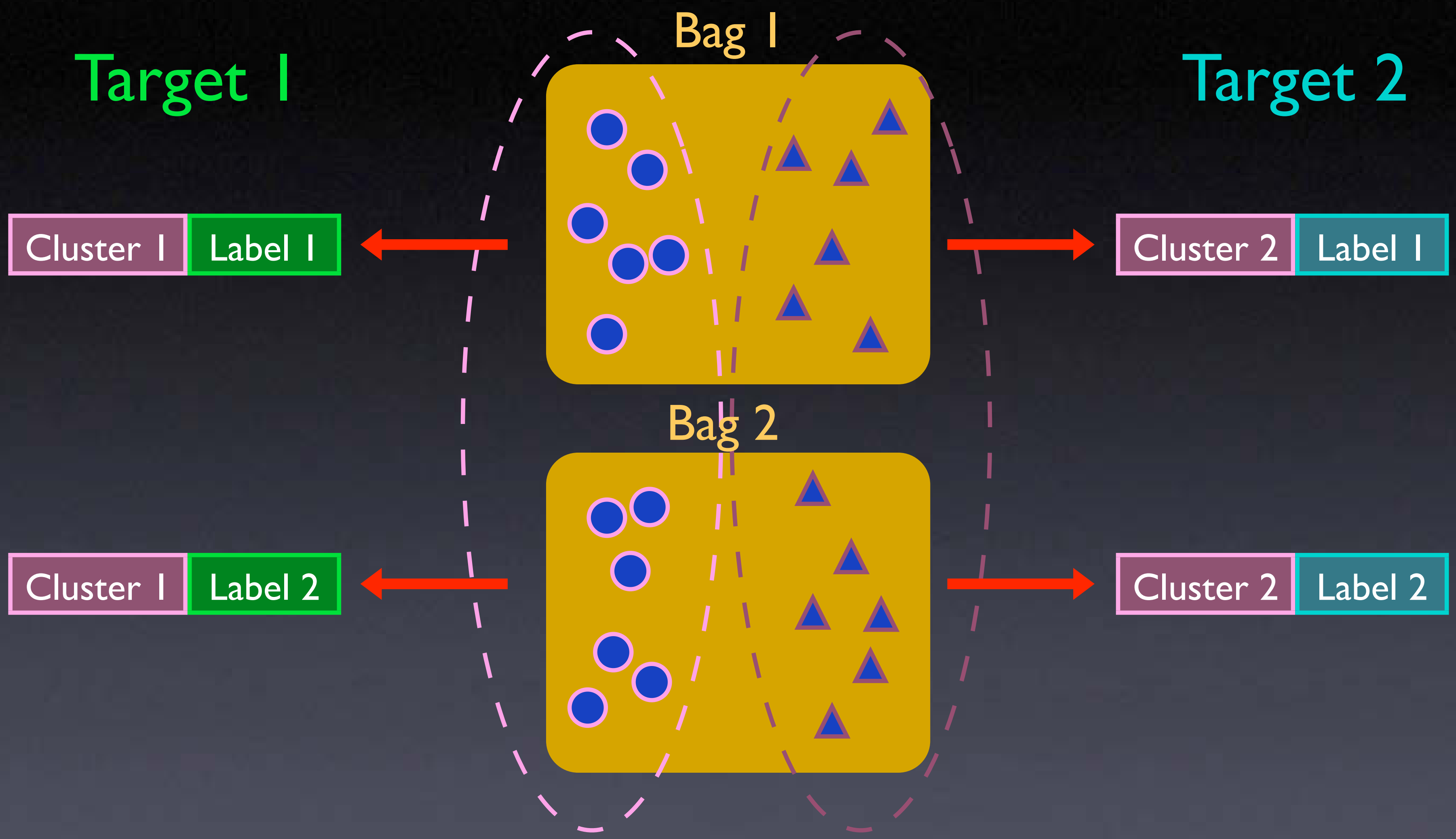


MIL with Structured Bags



Goal: infer cluster memberships to enable label prediction

MIL with Structured Bags



Goal: infer cluster memberships to enable label prediction

MI-ClusterRegress

1. Cluster entire collection of data into k clusters

Mixture model $$f(x_i) = \sum_{c=1}^{k} \alpha_c f_c(x_i) \qquad f_c(x) = \mathcal{N}(M_c, \Sigma_c)$$ Gaussian

2. Create weighted exemplar for bag B, cluster c

Membership prob. $$p(c|x_i) = \frac{\alpha_c p_c(x_i|M_c, \Sigma_c)}{p(x_i)} \qquad w_{cB} = \frac{1}{|B|} \sum_{i=1}^{|B|} p(c|x_i) x_i$$ Exemplar

3. Build k regression models
   - Model L_c: map all bag exemplars w_{cB} to bag labels
4. Select the regression model $L_{c'}$ that best fits the labels

MI-ClusterPredict

- Predicting the label of a new bag B':

1. Classify items in B' into the k clusters

Membership prob.
$$p(c|x_i) = \frac{\alpha_c p_c(x_i|M_c, \Sigma_c)}{p(x_i)}$$

2. Create an exemplar for the items in cluster c'

Exemplar
$$w_{c'B'} = \frac{1}{|B'|} \sum_{i=1}^{|B'|} p(c'|x_i) x_i$$

3. Use $L_{c'}(w_{c'B'})$ to predict the bag's label

Crop Yield: Methods Evaluated

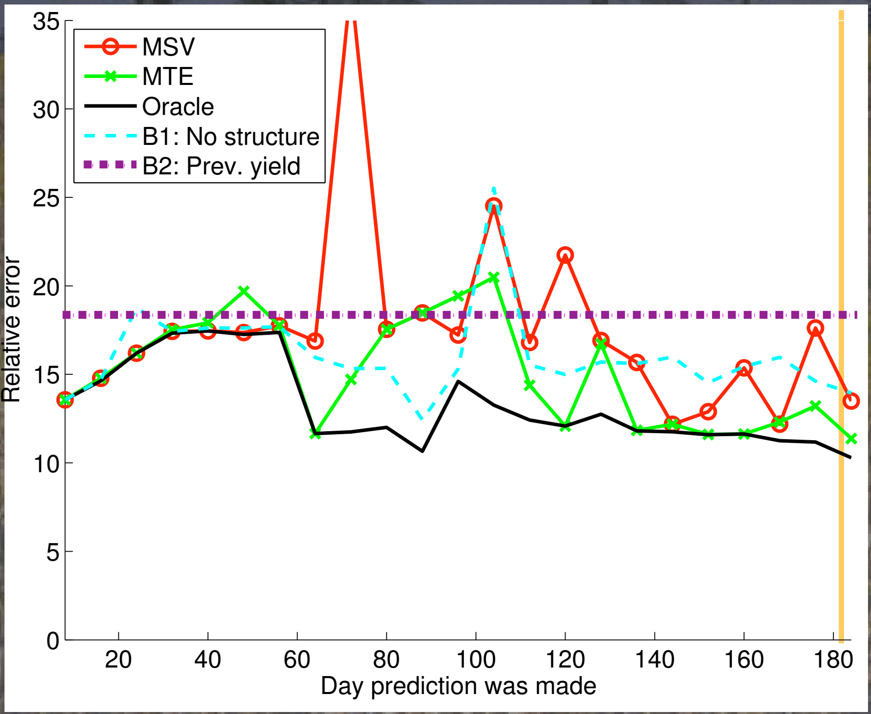
- MI-ClusterRegress Model Selection methods:
  - Complexity: minimum # of support vectors
  - Training: minimum error on training data
  - Oracle: minimum error on test data

- Baselines
  - B1: Exemplar = mean pixel (no structure)
  - B2: Last year's yield

Results: Train on '01-04, test '05

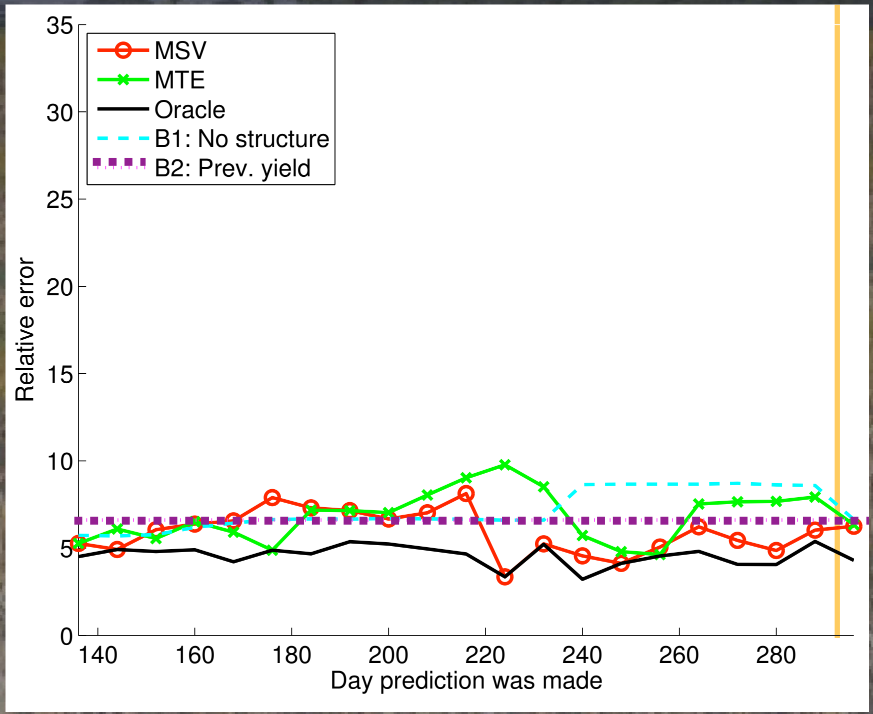
- CA: 42 counties, subsample 100 pixels/county
- Using K = 30 local models, select the best
- Same input data used to predict different crops

Harvest time



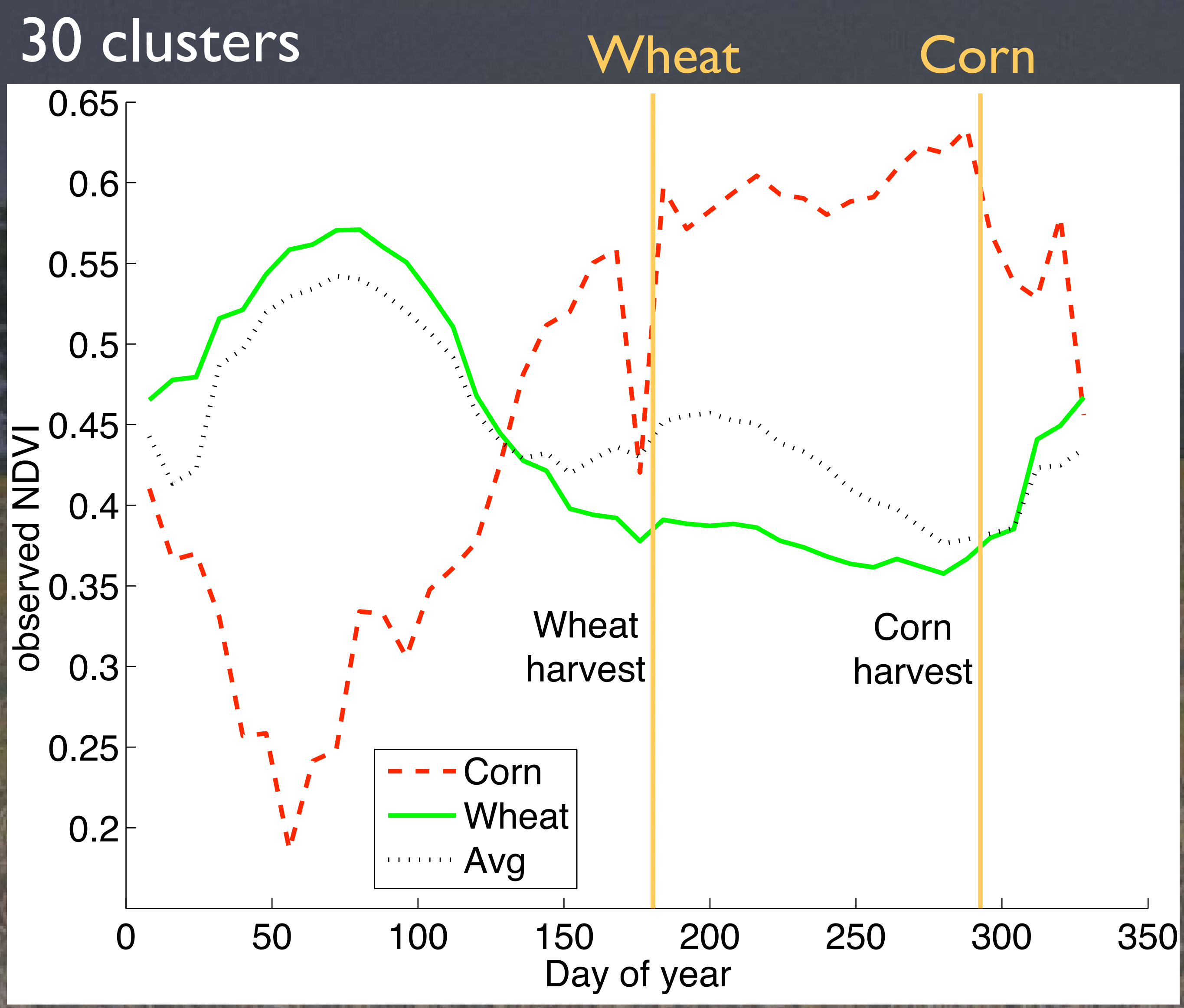
Wheat

Harvest time



Corn

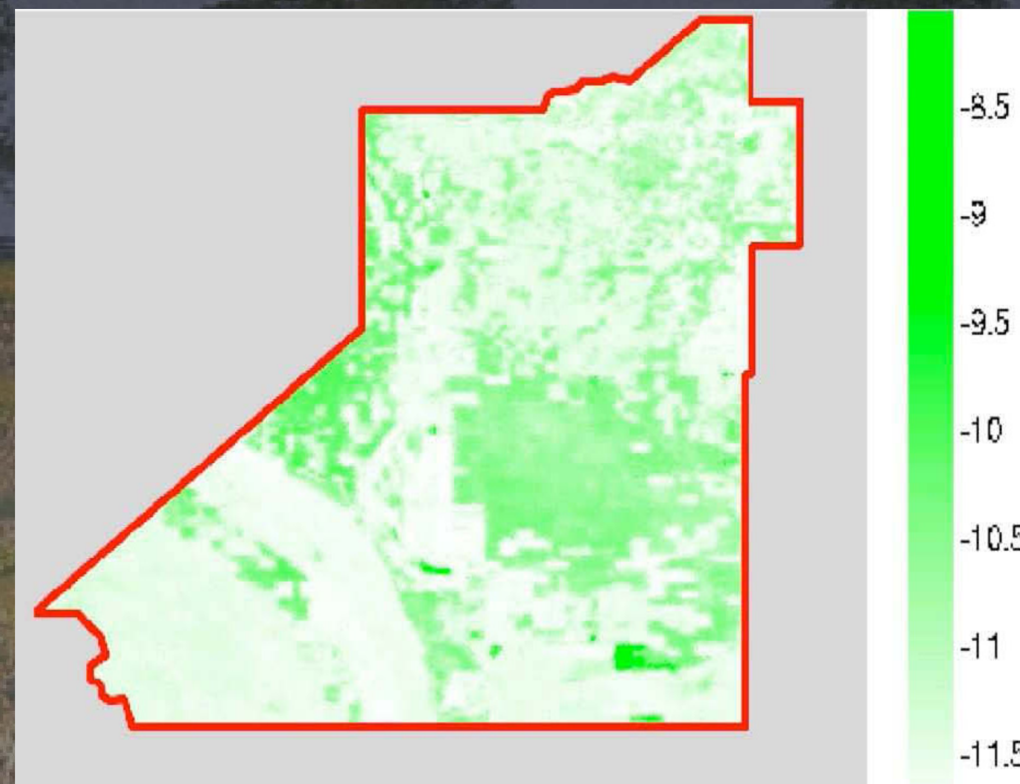
Model Selection: Clusters Chosen



Pixel Salience: Kings County, CA

Google Maps

Wheat Salience, Day 72



Conclusions and Future Work

- MIR with structured data: challenging new problem
- MI-ClusterRegress: Build per-cluster regression models that predict bag labels based on item relevance
- Crop yield prediction
    - 5-10% relative error in predictions 4 months before harvest
    - Bonus: item relevance provides per-crop maps
- Future work
    - Larger per-county samples, more crops, more counties
    - Other model selection heuristics
    - Relax Gaussian assumption on internal bag structure

Dr. Wagstaff's work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration, and supported by a grant from NASA's Applied Information Systems Technology Program. Dr. Lane's work was supported by NIMH grant number 1R01MH076282-01 from the CRCNS program. Their collaboration is also supported by NSF grant numbers 0325329 and 0705681.